The Emotional Voices Database: Towards Controlling the Emotion Dimension in Voice Generation Systems

Adaeze Adigwe[2⋆], Noé Tits[1⋆], Kevin El Haddad[1[0000−0003−1465−6273]], Sarah Ostadabbas[2], and Thierry Dutoit[1]

[1] numédiart Institute, University of Mons, 7000, Belgium
[2] Augmented Cognition Laboratory, Northeastern University, Boston, USA
{noe.tits, kevin.elhaddad, thierry.dutoit}@umons.ac.be
ostadabbas@ece.neu.edu

Abstract. We present a database of emotional speech intended to be open-sourced and used for synthesis and generation purpose. It contains data for male and female actors in English and a male actor in French. The actors were recorded in 2 different anechoic chambers. Each actor was asked to utter a subset of the sentences from the CMU-Arctic database for English speakers and from the SIWIS database for the French speaker. The database covers 5 classes (amusement, anger, disgust, sleepiness and neutral) with the goal to build, in the future, synthesis and voice transformation systems with the potential to control the emotional dimension in a continuous way. The number of sentences are shown in Table 1

Table 1: Number of utterances for each speaker and emotion

Speaker	Gender	Language	Neutral	Amused	Angry	Sleepy	Disgust
Spk-Je	Female	English	417	222	523	466	189
Spk-Bea	Female	English	373	309	317	520	347
Spk-Sa	Male	English	493	501	468	495	497
Spk-Jsh	Male	English	302	298	-	263	-
Spk-No	Male	French	317	-	273	-	-

In order to validate our database, we show the performance of the data in two experiments.
The first one involves a voice transformation system: a neutral voice is transformed into an emotional one. This system is built by extracting speech features with the WORLD vocoder of both source and target emotional voices, performing a Dynamic Time Warping to align the features in time and computing a regression between the source and target features via a simple Multi-Layer Perceptron (MLP) of 6 feedforward

⋆ These authors contributed equally to this work

hidden layers in which each hidden layer is constituted of 1024 hyperbolic tangent units.
The second experiment involves a Text-to-Speech synthesis system. In this experiment, we choose DCTTS, a system that seems to combine advantages of several systems. DCTTS models a sequence-to-sequence problem with a encoder-decoder structure along with an Attention Mechanism. Its architecture is entirely CNN-based, there is no RNN component. We investigate the adaptation this model to obtain an emotional TTS. To do this, we fine-tune system to the neutral voice of one of the actresses of in our database. We then fine-tune the obtained neutral TTS model with each emotion class of the same speaker. We evaluate the quality of the emotional speech synthesized through a MOS test for each emotion according to two criteria: confidence in the perception of the emotion specified and the intelligibility. (between 0 and 5). The results that are shown in Table 2 are encouraging.

Table 2: MOS test results of synthesized files

	Intelligibility	Confidence
Amused	2.01 ± 0.24	2.00 ± 0.27
Angry	2.76 ± 0.25	2.10 ± 0.28
Disgusted	2.17 ± 0.27	2.27 ± 0.30
Neutral	3.60 ± 0.26	3.59 ± 0.24
Sleepy	2.59 ± 0.28	3.29 ± 0.26

Currently, we are investigating the development of a multi-emotional TTS system with the possibility to control the intensity of emotional categories. We implemented a modified version of DCTTS that takes an encoding of the emotion category at the input of the decoder. During training, a simple one-hot encoding is used. But at synthesis stage, we can modify the intensity of an emotion category by inputting numbers smaller or greater than one. The results sound encouraging and we are planning to do some subjective evaluations to assess the system.